1. 研究

* 异构网络

本周了解了关于异构网络的内容，认真阅读了韩家炜在KDD'12上的演讲。他以DBLP数据为基础，先后提出了多个问题，问题的难度逐步增加，分别为：

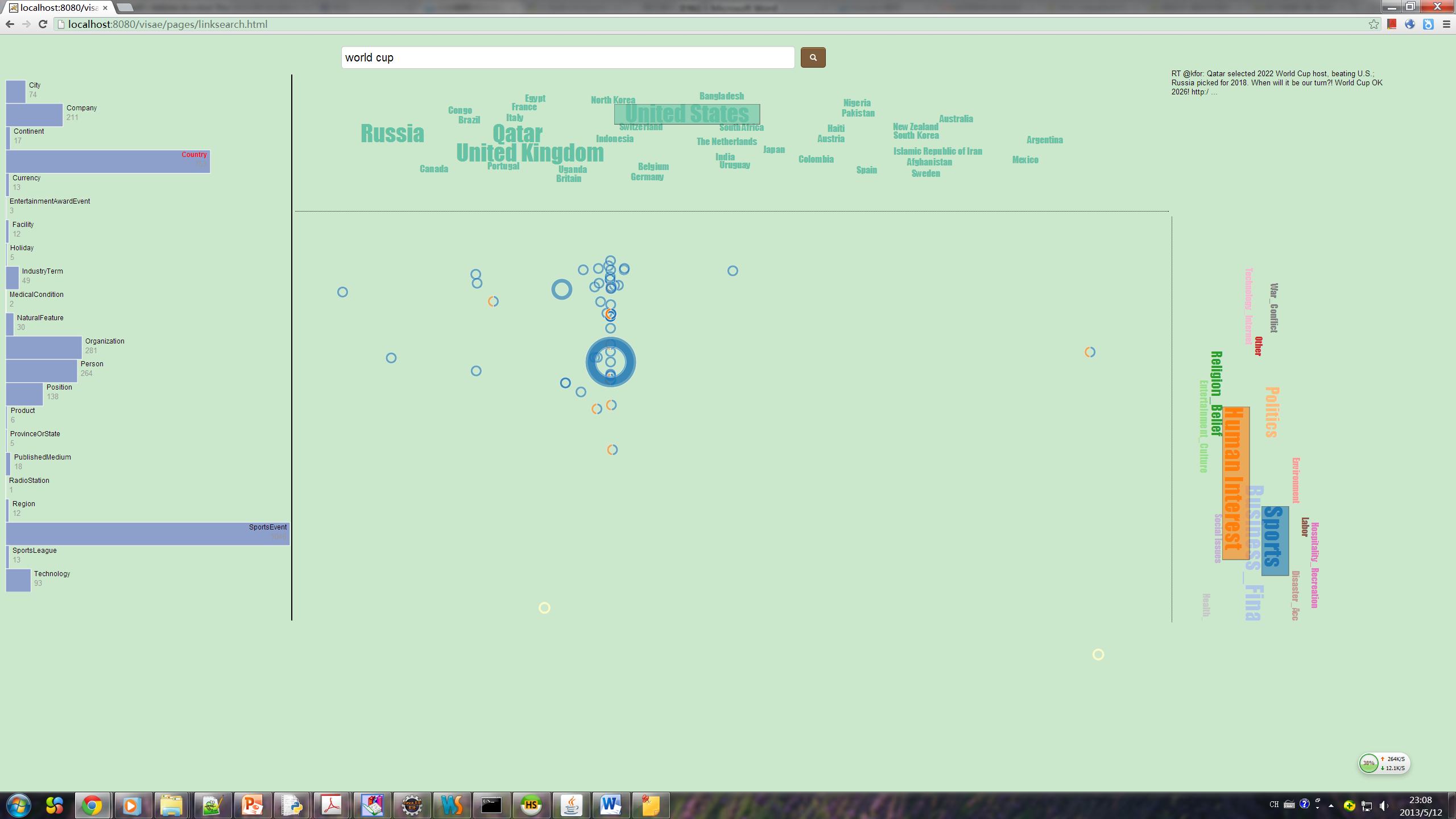


从作者发表的论文来看，目前已经解决了第五个问题，第六个问题正在解决过程中。与同学交流，一致认为在异构网络可视化中，可视化什么这是主要问题，如果仅仅是将网络通过点线图画出来，感觉没有什么新意，这个问题还需要进一步思考。

我将韩家炜的八个问题与我目前分析的Twitter数据对照，**发现Twitter数据其实也是异构网络数据**，其中存在话题、实体、用户、tweet，以及用户之间或tweet之间的联系，用户相互follow，tweet是否属于同一话题，我们解决的问题部分与similarity search相似。在搜索完之后，增加了可视化的部分从分类信息中获取想要的结果。

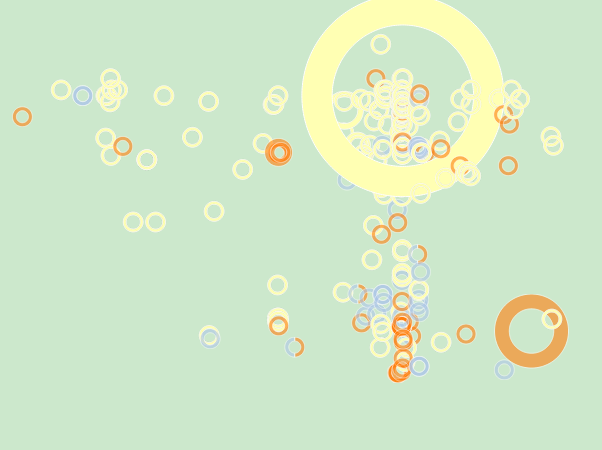
* Twitter可视化搜索

在原来的基础上重新对可视化的结果做了调整。柱状图是上周就改了的，本周主要是对布局做了调整，发现原来的计算结果中存在一个问题，在横向上节点大量重叠，这与预期的结果不相符。经过查找，发现是在计算过程中，有一些tweet的id有重复，删掉后基本与预期的一致。结果如下：



图中是查找”world cup”之后的结果，左侧列出了所有与world cup相关的命名实体的分类（总共39种，这里只有22种，其他相关tweet数量为0）。中间上部是选中”Country”类实体后，呈现的具体实体。选择”United States”高亮所有与之相关的tweets，再选择抽象话题”Sports”和”Human\_Interest”，在所有高亮的tweets中进行分类显示。圆环的颜色表示与哪个抽象话题相关，如果两个都相关则将圆环分成多块。点击某个圆环则在右上角显示具体的tweet内容。

目前对于这个结果还存在一点问题，就是当搜索结果很多时，仍存在大量的重叠，如下图所示。初步考虑使用弹力模型对节点做布局，并将相同类别的聚类，这个将在下周实验。



1. 学习

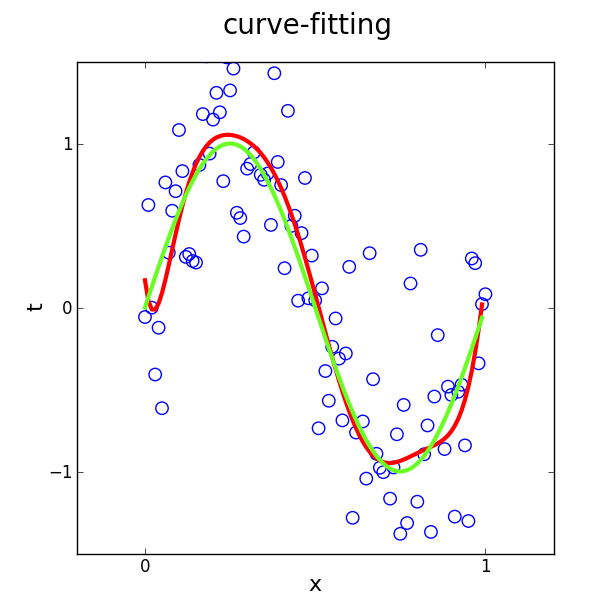
半年来，每天都在想着怎么去产生新的想法，一直没有认真学习可视化或相关的基础知识，如今停下来认真去读组内的教材，结合之前所犯的一些错误，对可视化也有了新的认识。例如，可视化元素由三方面组成：可视化空间、标记和视觉通道。标记用于表达定性信息，如分类，可以是任意的点、线、面、体。视觉通道则展现数据属性的定量信息，可以是空间位置、大小、饱和度/亮度。

同时，可视化不仅仅是画图，其实是一个包含多阶段的过程，从获取数据开始，对数据预处理、分析，再到可视化编码、设计，最后绘制。这是系统工程，任何一个阶段的问题都会对最终结果产生影响。学习可视化，数据分析也是要必须掌握的基本技能。

教材中，映像最深的格式塔理论，总结了感知的重要现象，是可视化设计的基本原则，这些原则对于设计是非常重要的，而我以前根本就没去思考过这方面的问题，没有全盘考虑过如何去表达信息。

在学习层次、网络数据可视化时，也了解了一些以前没有注意到的一些问题，例如，treemap的具体算法、点线图实现层次结构的具体过程。基本算法没有掌握，每次都是直接调用已有的模块，这对于创新来说是非常受限制的。这也限制了我的思维模式，总是拿来主义，而没有深入思考，不愿创新，这也正是工程思维和学术思维的差异所在了。以后要坚决抵制这种拿来主义的惯性思维。

本周还抽空做了一下应用数学的习题，长期与数学隔离导致基本的概念都已经不记得了，还好，经过回忆还是实现了用最小二乘法实现最大似然估计解决曲线拟合问题，如下图。



以后这种基本训练还是要坚持的，日积月累将对自己的学术研究有很大帮助。

1. 下周任务

* 再去看看Interactive Data Visualization，这本书也很经典，学习其中的一些经典可视化算法。
* 准备计算机前沿课程的报告。
* 进一步修改Twitter可视化搜索的结果。